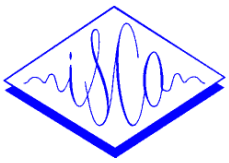


3rd European Conference on
Speech Communication and Technology
EUROSPEECH'93
Berlin, Germany, September 19-23, 1993

NEXT UTTERANCE PREDICTION BASED ON TWO KINDS OF DIALOG MODELS

Yoichi Yamashita and Riichiro Mizoguchi

The Institute of Scientific and Industrial Research, Osaka University,
8-1, Mihogaoka, Ibaraki-shi, Osaka, 567 Japan.
E-mail: {yama,miz}@ei.sanken.osaka-u.ac.jp

ABSTRACT

This paper describes a method of predicting user's next utterances in spoken dialog based on two kinds of dialog models, SR-plan and TPN. The SR-plan is a low level model of interactions composed of a stimulus and a response and the TPN is a high level model representing transitions of topics in dialogs. The dialog manager predicts next utterances and provides the language processing unit with templates of case frames which are associated with SR-plans and instantiated according to the preceding utterances and the topic information. An experiment shows that the utterance prediction improves the performance of the utterance recognition and drastically reduces the search space in terms of candidates in the input word lattice.

Keywords: Dialog model, Utterance prediction, Spoken dialog understanding

1. INTRODUCTION

The spoken dialog understanding system is one of the ultimate goals of speech research. The integration of techniques of signal processing and pattern recognition based on the bottom-up processing is not sufficient for realizing the spoken dialog understanding system. High level knowledge, such as knowledge about linguistics, background domain, dialog, and so on, is required for compensating the incompleteness of speech recognition. The knowledge about dialog is particularly useful for understanding the cooperative and goal-oriented spoken dialog because it proceeds under some constraints of dialog [1][2]; that is, participants of dialog neither ignore the opponent's utterances nor suddenly change the topic into a non-related one to it. A way of efficient use of dialog knowledge is next utterance prediction which provides the language processing unit with some useful constraints on meanings and a vocabulary in the utterance. We have already proposed a mechanism for predicting the next utterance of the user in computer-human interaction dialog based on utterance pair, which is modeled in terms of the SR-plans [3]. The SR-plan is a low level model capturing basic structure of dialog. In this paper another dialog model, TPN (topic packet network), which is a high level model representing topic transitions in dialogs is introduced in order to improve performance of prediction. Thus, two levels of dialog model are integrated in a mechanism for predicting the user's next utterances.

2. DIALOG MODEL

We have been developing a general speech interface system which is independent of the dialog domain, aiming at speech communication with various intelligent performance systems (IPS) [4]. Fig.1 shows the block diagram of the speech interface system. In order to keep the portability, the dialog manager in the interface system is designed so as to use as less domain-dependent knowledge as possible. In general, we can find at least two kinds of structures in dialog: one is utterance pairs and another is discourse segments [5]. In our dialog manager, named MASCOTS, the former is modeled by the SR-plan and the later by the TPN.

2.1. SR-plan

In a goal-oriented dialog, utterances are roughly classified into two types: a stimulus (requirement) utterance making a question, asking something to do, and so on, and a response utterance responding to the stimulus. That is, the goal-oriented dialog consists of utterance pairs whose components are the stimulus and the response to it. To grasp the characteristics of the dialog structure, we introduced the concept of SR-plan which represents an interaction of stimulus and response.

There are two types of SR-plans: one is the system SR-plan which is for dealing with an interaction constructed by an IPS's stimulus and a user's response, and the other is the user SR-plan for a user's stimulus and an IPS's response. We classified SR-plans into 17 categories according to the types of interactions. Fig.2 shows the structure of the SR-plan. An SR-plan basically

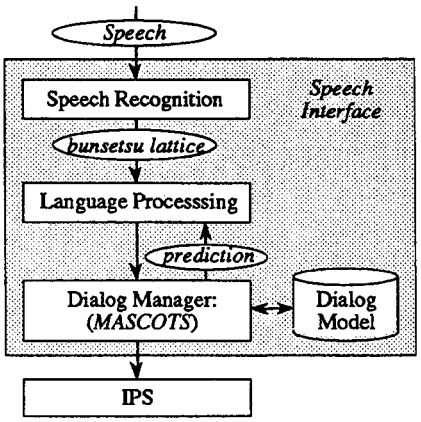


Fig.1 The block diagram of the speech interface system.

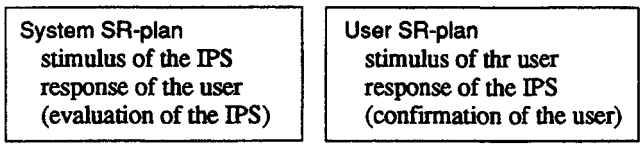


Fig.2 The structure of SR-plan.

consists of a sequence of stimulus and response, and a system and a user SR-plan can optionally contain an IPS's evaluation and a user's confirmation, respectively.

The prediction using the SR-plan is based on the dialog characteristics that the response comes after the stimuli. Hence, the SR-plan architecture is very powerful to prediction of user's next response because the response pattern can be easily associated with the preceding stimulus. However, it is difficult to predict the user's stimulus by using only the SR-plan because the SP-plan does not provide any constraints on meanings of stimulus utterances.

2.2. TPN

Topic transitions in a dialog can be viewed as a sort of hierarchy. In other words, the dialog begins with a rough topic and narrows down to more elaborated ones, and often comes back to the previous topic explicitly or implicitly. In addition, a topic can be elaborated into limited descendants. We believe that the relationship between a topic and its descendant topics is often

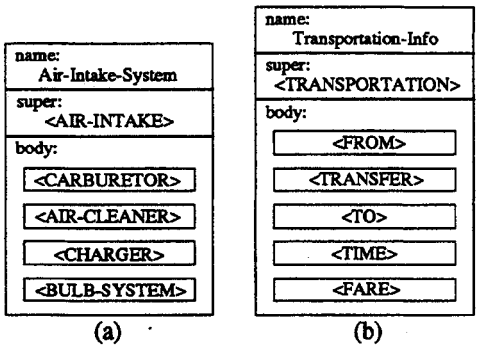


Fig.3 Examples of topic packets.

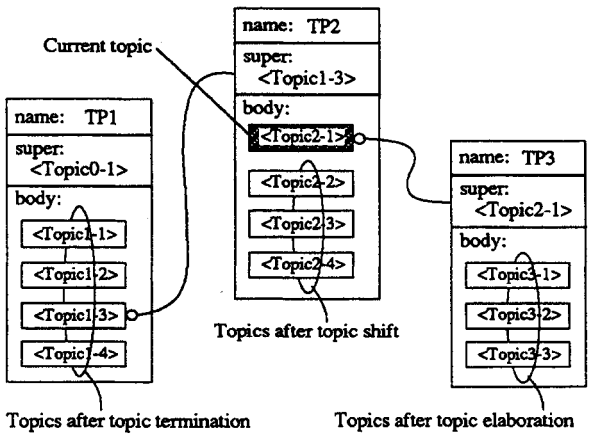


Fig.4 Predicted topics after three types of topic transitions.

independent of individual dialogs. We define such a set of descendants as a TP (topic packet). For example, in a dialog for designing a car, after the air intake becomes the current topic, the relevant topics may include a carburetor, an air cleaner, a charging mechanism, and a bulb system. The TP of <Air-Intake-System> shown in Fig.3 (a) explains such topic transitions. These topics may also appear in other dialogs for different purposes, such as the diagnosis of a car. The TP which represents a local topic transition is basically usable for several types of dialogs. Fig.3 (b) shows another example of the TP.

Topic transitions in a dialog are modeled as traversing in a network of TPs, which is composed of some TPs linked each other and is named TPN (topic packet network). Each TP has a super-topic. When a topic moves to an elaborated one as the dialog continues, another TP whose super-topic is the previous topic is activated and one of the topics in it becomes the current topic. In general, topic transitions can be classified into three types:

(1) topic shift,
(2) topic elaboration, and
(3) topic termination.

In the TPN, these transitions are easily represented as a move to another topic in the same TP, a move to the descendant TP, and a return to the ancestor TP, respectively. Fig.4 shows three types of topic transitions in the TPN. Of course, the topic may not change in the next utterance and this case is classified into topic continuation.

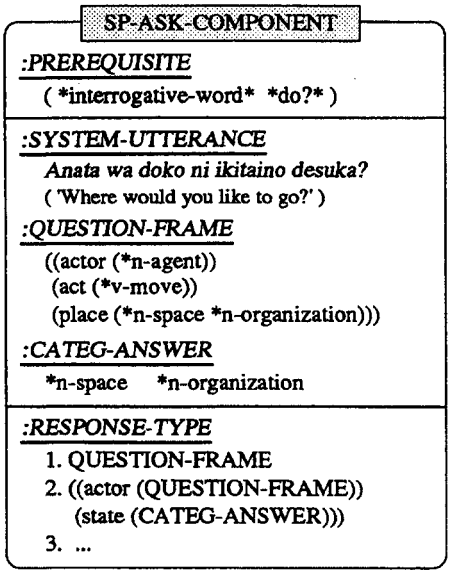
If topic transitions are described as a large network of each topic, it is too difficult to apply the network to another task of dialog. The introduction of topic packets increases the reusability of topic information because the local relationship between topics appears in many other dialogs. The TPN is introduced in order to compensate shortages of the SR-plan. Although the SR-plan does not have information on meanings, the TPN gives constraints on meaning of the user's next utterance by narrowing the range of potential topics in it.

3. MECHANISM OF UTTERANCE PREDICTION

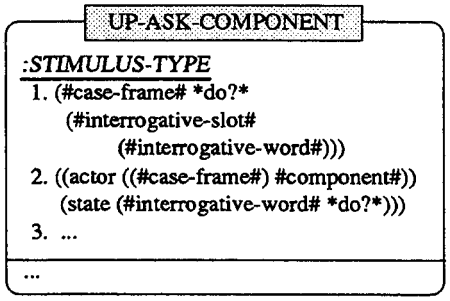
The mechanism of the utterance prediction is based on templates which are involved in SR-plans and instantiated according to the IPS's last utterance and the topic information.

3.1. Templates for Predicting Next Utterances

Each SR-plan has several slots as shown in Fig.5. The <RESPONSE-TYPE> slot in the system SR-plan is composed of some templates one of which is expected to match the user's next utterance. After <QUESTION-FRAME> was filled by analyzing the IPS's last utterance, which activates the system SR-plan, the templates in the <RESPONSE-TYPE> slot are instantiated. Since an activated system SR-plan is waiting for the user's answer, word categories of expected major points in the answer are described as <CATEG-ANSWER>. For example, the first and the second templates in the <RESPONSE-TYPE> slot in the SP-ASK-COMPONENT shown in Fig.5 (a) will match with the utterances such as '*Watashi wa Osaka-Daigaku he ikitai* (I'd like to go to Osaka University)' and '*Watashi ga ikitaino wa Osaka-Daigaku desu* (The place I'd like to go is Osaka University)', respectively, when the preceding IPS's utterance is '*Anata wa doko ni ikitaino desuka*? (Where would you



(a) System SR-plan



(b) User SR-plan

Fig.5 Prediction templates in SR-plans.

like to go?)'. Here, the word-category of '*n-organization' in <CATEG-ANSWER> can match with '*Osaka-Daigaku*'. Thus, in order to accept various expressions for the user's answer, the <RESPONSE-TYPE> slot should be described in terms of abstract templates.

The user SR-plan also has some templates in the <STIMULUS-TYPE> slot for predicting utterances. However, these templates can not be instantiated by using preceding IPS's utterances because the user SR-plan is activated by the user's utterance to be predicted. Therefore, the templates in the user SR-plan is instantiated based on the topic information. We classified topic-dependent information in the sentences into five kinds of *topic components* as follows.

(1) #case-frame#

Utterances under a particular topic contain limited number of words, especially verbs. Since proposition involved in the utterance is much dependent on the topic, this topic component expresses it in terms of the case frame of verbs. While the <QUESTION-FRAME> slot in the system SR-plan is a semantic representation for the IPS's utterance, the #case-frame# is associated with each topic and will match with the user's utterance.

(2) #component#

The #component# is used only in the user SR-plan UP-ASK-COMPONENT, in which the user asks objects or locations of an action. This is an abstract noun concept for such objects or locations.

```
1. ((actor (*n-human))
    (act (*v-getoff *do?*))
    (s-from (*n-transportation))
    (place (*where*)))
2. ((actor (((actor (*n-human))
                (act (*v-getoff))
                (s-from (*n-transportation))) *n-station))
    (state (*where* *do?*)))
3. ...
```

Fig.6 An example of prediction templates instantiated by the topic.

(3) #interrogative-words#

The #interrogative-words# represents what word replaces the #component# in the WH-questions. This is also used only in the UP-ASK-COMPONENT.

(4) #way#

The #way# corresponds to nouns representing a method and is used only in the UP-ASK-WAY in which the user asks how to do.

(5) #interrogative-slot#

The #interrogative-slot# indicates the case slot in the #case-fame# which contains #interrogative-words#.

The templates for prediction in the user SR-plan are formed with these topic components. How to instantiate topic components is a priori prepared for each topic. Fig.6 shows an example of instantiation of templates, which are shown in Fig.5 (b), when the topic is <TO>.

3.2. Constraints of topic

If there are activated system SR-plans, both the system and the user SR-plans are used for prediction. Otherwise, only user SR-plans are used because the user never answer in that case. A plausible topic in the next user's utterance is constrained by the TPN model. That is, the topic continuation or one of three types of topic transitions, mentioned in 2.2, is acceptable in the user's next utterance. Furthermore, when the IPS's last utterance is a stimulus and activates a system SR-plan, the topic shift and the topic termination can be removed from the prediction range of the next topic because they mean that the user ignores the IPS's stimuli.

All topics included in such a prediction range becomes candidates of the next topic. One of the candidates is selected as the most plausible topic using bottom-up information of the word lattice which is generated by the speech recognition unit. This process is carried out based on the scoring for each candidate topic according to the keyword matching with the word lattice. To this end, several keywords are associated with each topic. The language processing unit analyzes the word lattice using the templates instantiated according to the most plausible topic and the IPS's last utterance. If the analysis succeeds, it results in the identification of the topic and the type of the topic transition. Otherwise, the lattice analysis is repeated under the next plausible topic. When the user's utterance is a response, topic continues.

4. EVALUATION

We implemented a TPN shown in Fig.6 in the prototype system and carried out an experiment of utterance recognition with prediction on only 10 user's utterances in a small sample dialog which contains 20 utterances. The task of the dialog is how to get to the destination. The TPN was generated by hand based on the several simulated dialogs with the same task as that of the test dialog. The input word lattice is generated by simulation of speech recognition. We assume that the user speaks *bunsetsu* by *bunsetsu*, which is a Japanese syntactic unit.

Table 1 summarizes the recognition result of the user's utterances. In the recognition without the TPN, the templates for prediction consist of word categories representing many words which potentially appear in the task and many instantiated templates are generated namely due to various verbs. The introduction of the TPN model suppressed the template generation using only words associated to the assumed topic and lead to the improvement of recognition for two utterances (U6, U12).

Table 2 shows that constraints on dialog knowledge can prune the word lattice. Introduction ofboth the SR-plan and the TPN reduced more 40% of words in the lattice. When the topic is correctly identified, about 60% words can be reduced drastically.

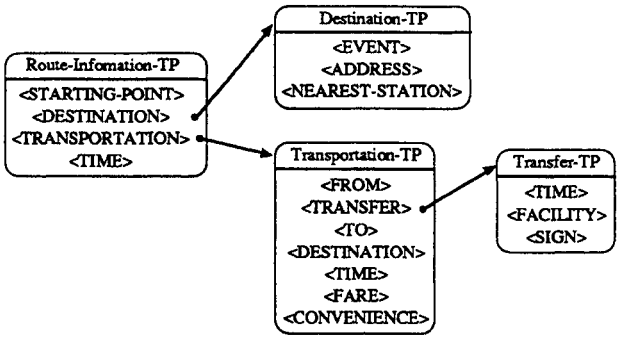


Fig.6 A TPN for dialogs on route information.

5. CONCLUSIONS

This paper describes a method of utterance prediction based on dialog models, which are based on two characteristics of dialogs: (1) a response comes after a stimuli, (2) topics do not suddenly change into non-related ones. We would like to describe the dialog knowledge in terms of domain-independent concepts in order to increase portability of the dialog managing system. The SR-plan architecture represents the basic characteristics of dialogs, that is utterance pairs, and is completely independent of the task of dialog. Each TP is a local transition pattern of topic in dialog and applies to various dialogs while the whole TPN is dependent on the task.

It is impossible to identify the topic in the utterance without the bottom-up information. In the prediction mechanism proposed, the most plausible topic is selected according to scoring result with keywords. How to recognize the topic will be a future work.

REFERENCES

[1] A.G.Hauptmann, et al.: "Using Dialog-Level Knowledge Source to Improve Speech Recognition", Proc. of AAAI '88, pp.729-733 (1988).

[2] A.K.Matrouf, et al.: "Adapting Probability-Transitions in DP Matching Process for an Oral Task-Oriented Dialogue", Proc. of ICASSP '90, pp.569-572 (1990).

[3] Y.Yamamoto, et al.: "Dialog Management System MASCOTS in Speech Understanding System", Proc. of ICSLP '92, Kobe, pp.1301-1304 (1992).

[4] Y.Yamashita, et al: "MASCOTS II: A Dialog Manager in General Interface for Speech Input and Output", IEICE Trans., E76-D, 1, pp.74-83 (1993).

[5] B.J.Grosz, et al.: "Attention, Intentions, and the Structure of Discourse", Comp. Linguist., 12, 3, pp.175-204 (1986).

Table 1 The result of utteranse recognition.

prediction methods	User's Utterance (utterance type)	U2 Res.	U4 Res.	U6 Sti.	U8 Sti.	U10 Sti.	U12 Sti.	U14 Sti.	U16 Res.	U18 Sti.	U20 Res.	Average
only SR-plan	instantiated templates	6	5	10	7	12	12	13	5	12	11	9.3
	rank of correct templates	2	1	5	1	9	4	4	2	2	1	3.1
	recognition result	○	○	×	○	△	×	○	○	○	○	8/10
SR-plan and TPN	instantiated templates	3	3	6	6	6	8	5	4	6	6	5.3
	rank of correct templates	2	1	2	1	3	7	1	2	1	1	2.1
	recognition result	○	○	○	○	△	○	○	○	○	○	10/10
rank of correct topic		—	—	1	1	1	2	1	—	1	—	1.2

(△: Sematically correct)

Table 2 The effect of reduction of word candidates per bunsetsu in the input lattice.

User's utterance (utterance type)	U2 Res.	U4 Res.	U6 Sti.	U8 Sti.	U10 Sti.	U12 Sti.	U14 Sti.	U16 Res.	U18 Sti.	U20 Res.	Average
No constraints	16.0	2.0	8.8	7.3	28.0	17.7	7.0	10.0	9.5	12.0	11.3
Using only SR-plan	14.0	1.0	3.8	4.0	21.5	14.7	5.3	5.0	7.8	7.0	7.8 (69.3%)
Using all topic candidates	6.0	1.0	5.2	4.8	15.5	12.0	3.0	9.0	4.5	6.0	6.4 (57.0%)
Using correct topic	5.0	1.0	3.4	3.5	8.5	8.0	2.7	4.0	4.0	5.0	4.5 (39.6%)