Model of Dialogue –Spoken Dialogue Recognition and Synthesis Based on Dialogue Model–

Riichiro MIZOGUCHI and Yoichi YAMASHITA
The Institute of Scientific and Industrial Research, Osaka University
8-1, Mihogaoka, Ibaraki, Osaka, 567 JAPAN
e-mail: miz@ei.sanken.osaka-u.ac.jp

1 Introduction

A number of problems must be solved before spoken dialogue systems are realized. Some problems are inherent in the dialogue, such as identification or complementing of neglected words, interpretation of the user's intention, and so on. There also exist problems which are inherent in speech interaction and do not appear in other media like written language. They are, for example, prediction of the next utterance, recovery from the unavoidable errors in speech recognition and listening of synthetic speech, and generation of prosodic features of the output speech. Since those problems are independent of the problem solver and inherent in dialogue and speech, they should be dealt with in the interface system with the dialogue management, and not in the internal function of the individual problem solver. From such a viewpoint, we have discussed them in this study, focusing on the utterance prediction for the spoken dialogue recognition and the spoken language generation appropriate to the dialogue context.

2 Dialogue Model for Topic Prediction

2.1 Topic prediction based on stochastic data

Speech recognition can be defined as a search problem in the word space under acoustical, syntactical, semantical, and pragmatical constraints. It is very important to reduce word candidates and prune the space by heuristic or stochastic knowledge. Prediction of user's next utterances is a useful technique to understanding of cooperated spoken dialogue, especially with the middle or large size vocabulary. Various kinds of dialogue knowledge, such as utterance pairs, topics, and so on, are indispensable for prediction of the utterances. From such a viewpoint, we have proposed a basic mechanism for utterance prediction. It was based on a topic transition model, named TPN (topic packet network). TPN is a static model of topic transitions represented as a sort of network. In a TPN, topic packets (TP) bundle some topics and they are a priori linked each other.

Our basic idea of the utterance prediction is arrangement of utterance patterns based on utterance pairs and instantiation of the patterns according to predicted topics. The prediction/identification of topic of the next utterance is one of the most important issues in our prediction mechanism. We discussed a method of predicting the topic from ambiguous utterance information, given as bunsetsu lattice. It predicts the topic of the utterance using top-down and bottom-up information, which are represented in terms of TPN and correlation of the word category and the topic, respectively. The former is extracted from a dialogue corpus by hand. The latter is automatically trained for each topic by using a topic-labeled dialogue corpus. The rank of the correct topic among 21 topic candidates was 3.7 and 2.6 for the bunsetsu lattice and the correct sentence, respectively, only with bottom-up information.

Furthermore, by using a topic transition model, these score were improved to 3.0 and 2.1, respectively. Future work includes the abstraction level of word categories and selective use of correlation of the word category and the topic.

2.2 Topic prediction based on utterance motivation model

Although TPN is a powerful mechanism for predicting the next topic, the topic of the utterance dynamically changes in a dialogue as the dialogue goes on. It is not easy to describe all the possible topic transitions in a static network especially for dialogue tasks with large vocabulary. A dynamic mechanism is appropriate for modeling topic transitions. We also discussed a new mechanism for enumerating possible topics of the next utterance by introducing a model of utterance motivation. In a goal-oriented dialogue, we carry out a dialogue in order to exchange focused information. Goals are achieved step by step through repeated interactions. We usually have a definite motivation of making an utterance in such a dialogue. This motivation of the utterance is closely related to the meaning of the utterance. A model of the utterance motivation is expected to flexibly predict topics in a dialogue.

We analyzed several dialogues concerning a task of route direction and trip reservation in order to investigate motivations of the utterance. Since it is reasonable to assume that a stimulus and a response in an utterance pair have the same topic, only the stimuli were analyzed. For example, "to make a comparison because an attribute has multiple values" is a motivation. Motivations could be divided into two different levels: communication and problem solving. The motivation of communication is triggered by the state of information exchange. A state that "an attribute has multiple values" is classified into the motivations at this level. The motivation of problem solving is related to how to use the derived information in the process of problem solving. "To make a comparison" is an example of the motivations of problem solving.

The motivations of communication are classified into 8 categories as listed in Table 1. Knowledge about the dialogue domain is necessary for the dialogue manager to know transferred information in a dialogue. The dialogue manager predicts the motivations of communication based on the state of transmission for each piece of information. We introduced two kinds of information packets to organize the domain knowledge. The "action sequence" represents a sequence of actions which are executed in order to achieve a goal. The "object" describes an assembly of information except for action sequences. The motivations of problem solving are classified into 10 categories as shown in Table 2. The purpose in the problem solving process by the user invokes the motivation at this level. A user modeling is a very important technique to predict the motivation of problem solving. This is a future work. The motivation of an utterance is modeled by combination of two levels of the motivation mentioned above.

A mechanism for predicting topics in the next utterance is described according to each utterance motivation. Assume that a slot of an information packet contains multiple values and "to compare" is predicted as the motivation of problem solving. The motivation "MC-c" is predicted based on the state of communication. Thus, the combination "MC-c" and "MP-A" is identified as the motivation of the next utterance. The topic prediction pattern for the motivation "MC-c&MP-A" is to enumerate topics in the information packet subordinated by the slot which contains multiple values. We carried out a small experiment of the topic prediction with and without the proposed motivation model. The motivation model could widely cover the possible topics of the next utterance. However, it tends to enumerate a lot

Table 1 Motivations of communication

MC-a:	A value is unknown
MC-b :	A value is ambiguous
MC-c :	An attribute has multiple values
MC-d :	An attribute or an action is unknown
MC-e :	Attrbutes or actions are exhausted
MC-f :	An object or an action sequence is ambiguous
MC-g :	To confirm something uncertain
MC-h :	An object or an action sequence is unknown

Table 2 Motivations of problrm solving

MP-A:	To compare
MP-B:	To select
MP-C:	To sort
MP-D:	To know/inform the reason
MP-E:	To know/inform the condition
MP-F:	To know/inform the related information
MP-G:	To correct
MP-H:	To satisfy the constraints
MP-I:	To know/inform the goal
MP-J:	To know/inform the completion

of topics at the end of the dialogue.

3 Context Dependencies of the Utterance in Dialogue

Utterances in spoken dialogue are produced with varieties of the surface sentence and the prosodic characteristics according to the dialogue context, even if they have the same meaning. Generation of the utterance appropriate to the dialogue context is a crucial issue of spoken dialogue synthesis. In this study, we analyzed the context dependencies of linguistic and prosodic characteristics for utterances in dialogue.

3.1 Surface sentences

In the frame work of speech output based on a concept-to-speech conversion system, the problem solver determines what to say using concept representation and does not care the surface expression and prosodic characteristics. It is essential to take account of the dialogue context for sentence generation in order to synthesize dialogue speech with high quality. The general speech interface modifies the concept representation according to the dialogue context and converts them into synthetic speech. The dialogue manager in the interface has roles of determining the surface sentence of the output message.

We investigated the difference between the sentence expressions generated with and without the dialogue context through an experiment of sentence generation from concept representations for two dialogue scenarios. Differences between the two sets of the sentences were categorized into 11 factors, such as insertion of conjunctions or other words, replacement of function words, ellipses, change of the word order, and so on. However all the factors are not necessarily preferred for sentence generation. Another experiment was carried out in order to investigate preference of factors of the sentence expression. Subjects were asked to select the most preferable one among several sentences provided with the dialogue context. As a result, many kinds of factors got high scores over 80%. Such preferred factors should be incorporated into sentence generation. We propose ideas of generating sentences with preferred factors using the dialogue context modeled in terms of the topic transition model and information transmission acts. Some preferred factors, however, are difficult to be dealt only with the knowledge independent of the dialogue task. In such a case, we must reconsider the information transmitted from the problem solver to the general speech interface.

3.2 Prosody

Prosodic characteristics of utterances in a dialogue vary in accordance with the dialogue context even if their surface sentences are the same. Many studies have been conducted for qualitative intonational prediction from dialogue information, especially for placement of the pitch accent for English sentences. However, the quantitative prediction is very important as well as the qualitative one, especially for F0 contour of Japanese sentences.

We proposed a method of modeling contextual effects on prosodic parameters in dialogue based on the two-step prediction. In this framework, the prosodic parameter is predicted by two sets of rules; T-rule and D-rule. The fundamental behavior of the parameters based on the syntactic features is captured by the T-rule, which is modeled for isolated utterances. Then, the D-rule adjusts the parameter to dialogue context using higher level information. The D-rules are derived by minimizing the difference between F0 parameters of the isolated and dialogue utterances. These rules predicted F0 local peak in the dialogue utterance focusing on the Japanese noun phrases containing proper nouns such as person's name or the day of the week in the scheduling task.

The T-rule was derived from 503 isolated sentences uttered by one speaker based on 7 syntactic features. The dialogue utterances were collected through 15 simulated dialogues. The utterances only of answerers were used for the D-rule modeling. They were uttered by 5 speakers which did not include the former speaker. We described high level information of the D-rule in terms of 5 features: the relationship to the preceding utterance, the existence of coordinate phrases, the position of the phrase in the utterance, the existence of the preceding filler, and the negligibility of the phrase. The D-rules were separately derived for the person's name and the day of the week. The D-rules reduced the prediction error from 18.9Hz to 11.2Hz and from 19.5Hz to 13.6Hz for the person's name and the day of the week, respectively. The relationship to the preceding utterance is a very important feature to predict the F0 local peak in dialogue. The derived D-rules for two categories are very similar to each other.

4 Conclusion

We proposed a new model of utterance motivation for predicting the topic of the next utterance. The evaluation experiment showed that the former mechanism based on the TPN model had a powerful predictive ability of the next topic. However, it has poor flexibility to dynamic change of dialogues. The utterance motivation model enables utterance prediction to be adaptive to situations in a dialogue.

The relationship of the linguistic and prosodic characteristics of the utterance and the dialogue context were also investigated in this study. The variety of the surface sentences were classified into 11 categories. Furthermore, two sets of the rules derived from isolated and dialogues utterances predicted very well F0 peak in dialogue.

Keywords: dialogue model, speech interface, dialogue context, topics, utterance prediction, prosody