Modeling of Man-Machine Dialogue

Shuji DOSHITA, Masahiro ARAKI and Tatsuya KAWAHARA
Department of Information Science, Kyoto University
Sakyo-ku, Kyoto 606-01, JAPAN
e-mail: doshita@kuis.kyoto-u.ac.jp

1. Introduction

We approach dialogue modeling by (1) collecting and analyzing dialogue data, (2) trying to establish total evaluation method of dialogue system, and (3) constructing dialogue model form the view of (1) and (2).

2. Collecting and Analyzing Dialogue Data

We collected following three types of dialogue data: (a) Wizard of Oz, Human-Machine dialogue on group scheduling task, (b) Human-Human dialogue on group scheduling task, and (c) Human-Human e-mail dialogue on route finding task.

For type (a), we collect 25 dialogues by 5 subjects. System utterance amounts to 655 and user utterance 360. We transcribe all data and add utterance type tag and intention tag in order to get statistical data.

For type (b), we collect same amount of data and analyze them concentrating on the difference between data (a). Then we get two qualitative observation; (1) blanks between turns are likely to be long in Human-Machine case (2) units of utterance are smaller than sentence in Human-Human case

For type (c), we collect 12 dialogues by 6 subjects. These data are used for case base in case-based parsing.

3. Proposal of Flexibility Ranking of Spoken Dialogue Systems and Total Evaluation Method

3.1 Evaluation of language processing subsystem

Recently, several research organizations developed spoken dialogue systems. Typical evaluation method of these systems are checking whether their output database records are correctly responding to input queries.But, there are some objections to this evaluation method. Because this method includes operations of application domain, general applicability of language processing subsystem is not evaluated.

Then an attempt to evaluate task independent language processing begins by setting task independent semantic representation as a goal of language processing. This is called SemEval. In SemEval, task independent semantic representation is constructed by predicate argument structure.

However, we think another point must be considered in spoken dialogue system. In spoken language, the system must deal with illegal input, that is to say, spontaneous speech. At spontaneous speech processing, the system cannot prepare model of all the illegal phenomena. Then the distance from input utterance to target semantic representation should become a measure of the difficulty of processing.

Therefore, we need to rank the utterance understanding based on such distance. It depends on the structural complexity of predicate argument structure. In surface expression, it almost correspond to flexibility of utterance. This is because we set a rank of flexibility of utterance. This attempt gives another point of view to spontaneous speech understanding, that is discussed by a difficulty of linguistic phenomena, such as interjections, missing function word, wrong word order etc. We use this structural complexity of predicate argument structure to evaluation of language processing subsystem.

In addition, complexity of task domain has an influence on language processing. In some simple task, if dialogue constraints are strong, expectation of next utterance can powerfully aid language processing. Then we regard flexibility of dialogue as complexity of dialogue model.

3.2 Flexibility of Utterance and Flexibility of Dialogue

In order to make the ability and generality of elementary subsystem apparent, we propose a ranking of flexibility of utterance and flexibility of dialogue. The ranking of flexibility of utterance is based on the structural complexity of semantic representation. On the other hand, Dialogue flexibility is based on the complexity of dialogue model. Table 1 shows the flexibility of utterance and Table 2 shows the flexibility of dialogue

Table 1: Flexibility of utterance

rank	definition	processing
ordered word sequence	content words and order can decide one semantic representation	template matching
simple sentence	semantic representation have only one predicate	parsing
complex sentence	value of argument can be predicate argument structure	recursive parsing

Table 2: Flexibility of dialogue

rank	definition	merit
automaton	flexibility of dialogue is restricted in regular grammar	easy to predict
plan recognition	flexibility of dialogue is restricted in CFG	easy to model the task
different knowledge	dialogue in different knowledge	model of user's mental state

Here we introduce **number of semantic hypothesis (NSH)** as an auxiliary criteria of sentence ranks. NSH aims at a semantic version of perplexity. NSH is calculated by the sum of number of frames. Each number of frames are calculated by the product of number of slots and number of instances.

3.3 Overall Evaluation of Dialogue System

In addition to above mentioned elementary evaluation, there is an attempt to evaluate overall dialogue system. Contrary to speech recognition evaluation, recorded date cannot be used in such attempted method. One way is to evaluate interactive process with human users. But once evaluation includes human factor, evaluation loses the objectivity. The other way of overall system evaluation is through system to system automatic dialogue. In Japan, system to system dialogue competition in minimum route search at subway is proposed. This is a competition of system's ability of problem solving and of conciseness of dialogue.

We extend this criteria of system to system dialogue. Because we want to establish evaluation method of spoken dialogue systems, then we decide to include linguistic random noise in the communication channel of the simulator. This noise simulates speech recognition errors. Then the point of this evaluation is how to repair or manage such misrecognized sentence by robust linguistic processor or dialogue management strategy.

In this method, system performance is evaluated by average achievement rate and average number of turns. But these values are dependent on task. Then, we must make clear the flexibility of utterance and flexibility of dialogue in each task by above mentioned method.

3.4 Example of Evaluation by Automatic Dialogue

We select minimum route search at subway as a task domain. The flexibility of utterance is limited to ordered word sequence rank and the flexibility of dialogue is automaton rank. The employed automaton has 8 states and the number of type of utterance is 9. Total words in this task are about 70 words, most of the words are name of the station in the map.

What we want to know by this evaluation is (1) how much speech recognition errors affect a length of dialogue, and (2) what type of dialogue strategy is useful in given accuracy.

In this experiment, we set error rate of speech recognition is 10 % and 30 %. We simulate speech recognition error by dropping one content word at given rate. This type of error reflects the errors occurred in case of template matching method.

Also, we prepare two dialogue strategy; ask back type and infer type. Ask back strategy means that if system can detect speech recognition error, then ask again by the utterance "*Mou ichido itte kudasai.* (I beg your pardon?)" On the other hand, infer type strategy means that system try to infer what was to be said if possible. Recovery strategy is that if predicate word drops, then system infers the predicate by the rest of the content words; if noun drops, then system give up recovering and ask again.

Multiplying these two conditions, we examine four types of dialogue. Three dialogue experiments are done for each type. All the cases (i.e. 12 cases) reach the goal because of simple rank of utterance and dialogue. Then we count a number of turns and calculate average turns to examine a system performance. Table 3 shows the result of experiment.

Table 3: Average turns under different dialogue conditions

error rate(%)	dialogue strategy	average turns
0	—	102.0
10	ask back	112.0
10	infer	112.7
30	ask back	148.7
30	infer	122.7

First column shows that it takes 102 turns to solve the problem if there is no recognition errors. It is a baseline of other experiment. Second and third columns show the result of examining the available of two types of dialogue strategy respectively, under 10 % recognition errors. It yields about 10 % of redundant interactions by recognition errors. In spite of different dialogue strategy, there is little difference in the average of turns. Fourth and fifth columns show the result under 30 % recognition errors. There is another 21.2 % of redundant interactions in ask back strategy.

By this result, in ordered word sequence rank and automaton rank, we can say that infer type strategy is not so effective at relatively low recognition errors. But if recognition errors is high, infer type strategy is more available than ask back strategy.

4. Cooperative Spoken Dialogue Model

Here, we propose a dialogue model which has two major characteristics. These are 'robust' and 'cooperative'. In order to realize these two important features, our model has two inference spaces. We define these spaces as *Conversational Space* (CS) and *Problem Solving Space* (PSS). CS is a kind of dynamic Bayesian network that represents a meaning of utterance and general dialogue rule. CS treats a 'robust' aspect. Also, PSS is a network so called *Event Hierarchy* that represents the structure of task domain problems (our domain is group scheduling). PSS treats a 'cooperative' aspect.

In constructing CS and making inference on PSS, system's process, from meaning understanding through response generation, is divided into five steps. These steps are (1) meaning understanding, (2) intention understanding, (3) communicative effect, (4) reaction generation, and (5) response generation. The relationship between two spaces and these five steps is shown in Fig. 1.

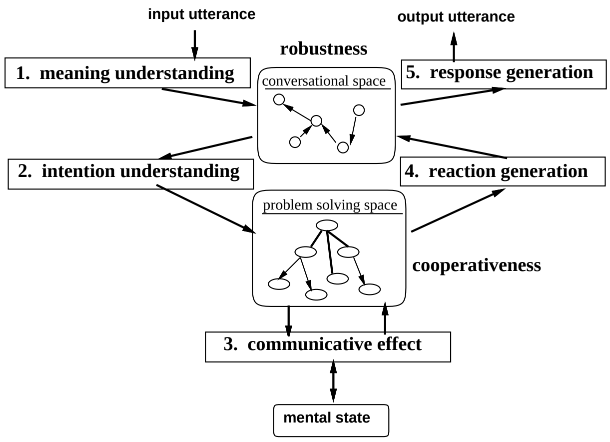

Figure 1: Five steps model of dialogue process

Meaning understanding step constructs CS and response generation step compose a surface expression of system's response from the part of CS. Intention understanding step make correspondence utterance type in CS with action in PSS. Reaction generation step selects a cooperative reaction in PSS and expands the reaction to utterance type of CS. The status of problem solving and declared user's preference are recorded in mental state by communicative effect step.Then, from our point of view, we regard cooperative problem solving dialogue as a process of constructing CS and achieving goal in PSS through these five steps.

5. Conclusion

In this report, we describe dialogue data collection, total evaluation method of dialogue system, and cooperative dialogue model.

Keywords: dialogue database, dialogue modeling, spontaneous speech