A METHOD FOR EVALUATING INCREMENTAL UTTERANCE UNDERSTANDING IN SPOKEN DIALOGUE SYSTEMS

Ryuichiro HIGASHINAKA, Noboru MIYAZAKI, Mikio NAKANO, Kiyoaki AIKAWA
NTT Communication Science Laboratories
NTT Corporation
3-1 Morinosato-Wakamiya, Atsugi, Kanagawa 243-0198, Japan
{rh, nmiya, nakano}@atom.brl.ntt.co.jp, aik@idea.brl.ntt.co.jp

1. Introduction

Evaluation measures for components in a spoken dialogue system :

- **Speech Recognizer → WER (word error rate)**
- **Speech Recognizer and Language Understanding Component → CER (concept error rate)**
- **Speech Recognizer, Language Understanding Component and Discourse Understanding Component → ??**

How can we evaluate the three components (understanding components) as a whole ?

Such an evaluation measure is especially needed, because we are promoting an understanding method ISSS (Incremental Sentence Sequence Search) which causes the dialogue state to update frequently.

We propose to create an evaluation measure by finding an equation that associates the behavior of the understanding components with the system's performance.

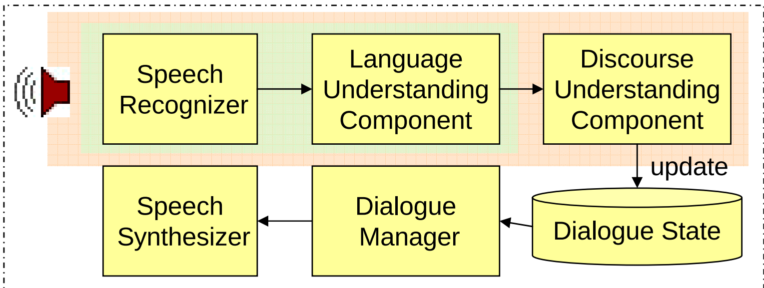

Fig1. Basic architecture of a spoken dialogue system

What's ISSS : **ISSS accepts both sentences and sentence fragments (i.e., words, phrases) and incrementally updates the dialogue state. If ambiguity is found in the understanding of the fragments, ISSS holds multiple dialogue states ordered by priority, so that the system can decide on a single dialogue state after any speech interval.**

2. Approach

Find the representation of the understanding components' behavior

We label the system's dialogue state (dialogue state hypothesis) in two respects.
(1) the correctness of the dialogue state itself (Fig2)
(2) the correctness of the update (Fig3)
We derive ten metrics to express the correctness of the dialogue states in a dialogue. (see 2.1)

Find the representation of the system's performance

We use task completion time. Task completion time correlates closely with user satisfaction.

Create an equation that can estimate the system's performance from the understanding components' behavior.

We perform a multiple linear regression analysis.

When a misrecognition happens, the error is normally inherited to the next dialogue state, thus it is not appropriate to use only the resulting dialogue state itself for its correctness. (Fig1)

How correct is this dialogue state?

System's Dialogue States

Slot	Initial Dialogue State	"from three"	"to four"
Date	--	--	--
Room	--	--	--
Start Time	--	2 o'clock	2 o'clock
End Time	--	--	4 o'clock

Misrecognition occurred and error inherited !

Correct Dialogue States

Slot	Initial Dialogue State	"from three"	"to four"
Date	--	--	--
Room	--	--	--
Start Time	--	3 o'clock	3 o'clock
End Time	--	--	4 o'clock

Fig1. Example of dialogue state updates

NOTE : We use a frame representation for dialogue states, each attribute-value pair is called a slot.

Slot	Dialogue State Hypothesis	Label	Dialogue State Reference
Date	5/20	correct	5/20
Room	--	deletion	Room 3
Start Time	2 o'clock	insertion	--
End Time	3 o'clock	substitution	4 o'clock

Fig2. Labeling of a dialogue state

Initial Dialogue State

Date	5/20
Room	--
Start Time	--
End Time	--

Slot	Dialogue State Hypothesis	Label	Dialogue State Reference
Date	5/20	correctly left	5/20
Room	--	update deletion	Room 3
Start Time	2 o'clock	update insertion	--
End Time	4 o'clock	correct update	4 o'clock

Fig3. Labeling of a dialogue state update

2.1 Ten Metrics

5 metrics concerning the dialogue state itself

$$\text{slot accuracy} = \frac{\#\text{ of correct slots}}{\#\text{ of slots}} \qquad \text{insertion error rate} = \frac{\#\text{ of insertion}}{\#\text{ of slots}}$$

$$\text{deletion error rate} = \frac{\#\text{ of deletion}}{\#\text{ of slots}} \qquad \text{substitution error rate} = \frac{\#\text{ of substitution}}{\#\text{ of slots}}$$

$$\text{slot error rate} = \frac{\text{total }\#\text{ of insertion, deletion, and substitution}}{\#\text{ of slots}}$$

4 metrics concerning the dialogue state update

$$\text{update precision} = \frac{\#\text{ of correctly changed slots}}{\#\text{ of changed slots in hypothesis}}$$

$$\text{update insertion error rate} = \frac{\#\text{ of changed slots in hypothesis}}{\#\text{ of unchanged slots in reference}}$$

$$\text{update deletion error rate} = \frac{\#\text{ of unchanged slots in hypothesis}}{\#\text{ of changed slots in reference}}$$

$$\text{update substitution error rate} = \frac{\#\text{ of incorrectly changed slots in hypothesis}}{\#\text{ of changed slots in reference}}$$

- **We calculate above nine metrics for each dialogue state, and use their average values for their values in a dialogue.**
- **We use additional one metric shown below.**

1 metric concerning all dialogue states in a dialogue

$$\text{speech understanding rate} = \frac{\#\text{ of dialogue states with 100\% slot accuracy}}{\#\text{ of dialogue states in a dialogue}}$$

3. Experiment

Domain	Meeting room reservation
Speech recognizer	Julius 3.1p
Speech synthesizer	FinalFluet (NTT SP Lab)
Number of subjects	18 (male : 9, female : 9)
Number of collected dialogues	180 (3595 utterances)
Task completion rate	63.6 %
Number of dialogues used for analysis	108
Number of dialogue strategies *	2
Number of task patterns *	5

Dialogue State Reference Tagging :

References are semi-automatically created using a simulation system and then manually corrected.

**Dialogue Strategy and Task Pattern :*

Task completion time is normalized using
(1) Dialogue Strategy (dialogue manager's behavior)
(2) Task Pattern (room reservation patterns)
to focus only on dialogue states and the system's performance.

4. Results

- **We performed a multiple linear regression analysis using**
  (1) the task completion time normalized by the task pattern and the dialogue strategy as the explained variable
  (2) the ten metrics (see 2.1) as explaining variables
- **By stepwise regression, seven metrics were incorporated as a result. Below is the equation :**

$$Y = -4.19 - 12.49x_1 + 12.77x_2 - 0.03x_3 - 17.74x_4 + 4.54x_5 + 2.11x_6 + 2.98x_7$$

Where Y is the predicted task completion time, x_1 the insertion error rate, x_2 the substitution error rate, x_3 the update precision, x_4 the update insertion error rate, x_5 the update deletion error rate, x_6 the update substitution error rate, and x_7 the speech understanding rate.

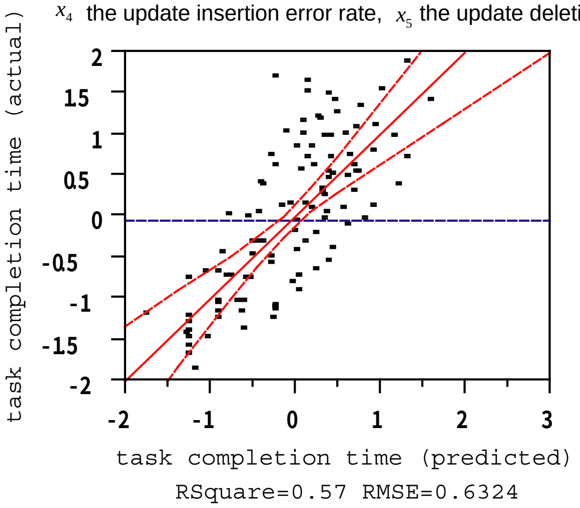

Fig4. Distribution of actual and predicted task completion times

	task completion time
slot accuracy	-0.40
insertion error rate	-0.07
deletion error rate	0.29
substitution error rate	0.40
slot error rate	0.40
update precision	-0.45
update insertion error rate	0.15
update deletion error rate	0.62
update substitution error rate	0.24
speech understanding rate	-0.42

Table1. Correlation coefficients of the ten metrics against the task completion time

As Fig4 shows, the model fits comparatively well with RSquare 0.57, the RSquare Adjusted 0.54, and RMSE (Root Mean Square Error) 0.63.

The correlation coefficients of the ten metrics against task completion time are shown in Table1.
The update deletion error rate has a relatively high correlation with correlation coefficient 0.62 followed by - 0.45 of update precision.

5. Summary and Future Plans

- **We proposed a method for evaluating incremental utterance understanding, which involves speech recognition, language understanding, and discourse processing in spoken dialogue systems, by performing a multiple linear regression analysis using the task completion time as the explained variable and various metrics concerning dialogue states as explaining variables.**
- **The resulting model shows a validity as an evaluation measure, and indicates that the use of both the dialogue states and their way of update is effective.**

Our future plans include: ***validation of our approach in other domains (e.g., more complex domains, more real-world-based domains), use of user satisfaction metrics in addition to task completion time.***