Evaluating Discourse Understanding in Spoken Dialogue Systems

Ryuichiro Higashinaka, Noboru Miyazaki, Mikio Nakano, and Kiyoaki Aikawa
NTT Communication Science Laboratories, NTT Corporation
Atsugi, Kanagawa 243-0198, Japan

Overview

- We propose a method for creating an evaluation measure for discourse understanding in spoken dialogue systems.
- We enumerated possible discourse-understanding-related metrics and used the metrics to create by regression methods a discourse understanding measure that correlates closely with system performance.
- The correctness of a dialogue state update is the most important factor in improving system performance.

Problem

- Discourse understanding is a process of updating a dialogue state by a user utterance.
- There is no appropriate measure to evaluate dialogue state sequences.
- For example, which of the two metrics below better evaluates the discourse understanding?

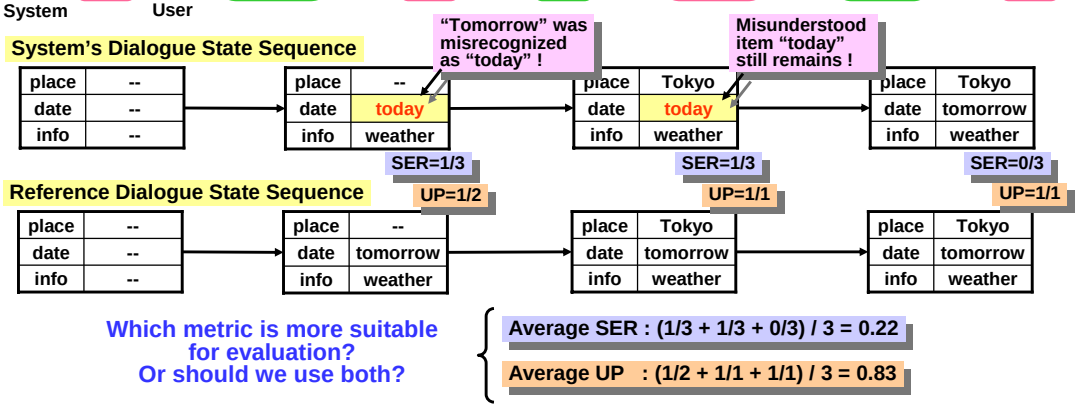
Metric 1

$$\text{Slot Error Rate (SER)} = \frac{\text{\# of incorrect slots}}{\text{\# of slots}}$$

Metric 2

$$\text{Updated Precision (UP)} = \frac{\text{\# of correctly updated slots}}{\text{\# of updated slots}}$$

-- Problem in an example dialogue --



Which metric is more suitable for evaluation? Or should we use both?

Average SER : (1/3 + 1/3 + 0/3) / 3 = 0.22

Average UP : (1/2 + 1/1 + 1/1) / 3 = 0.83

Approach

- Enumerate possible discourse-understanding-related metrics.
- Use the metrics to create by regression methods a discourse understanding measure that correlates closely with system performance.

Explaining variables	possible metrics about discourse understanding
Explained variables	Task completion time as system performance metric

Experiment: Data Collection

- Dialogue data collection in two domains.
  - Weather information service domain (WI)
  - Meeting room reservation domain (MR-1, MR-2)
- We hand-annotated reference dialogue states.
- We calculated the values of 13 metrics (see the table below) for each dialogue.
- Task completion times were normalized to make them comparable among systems.
- We used multiple linear regression and support vector regression as regression methods.

	WI	MR-1	MR-2
Subjects	12	28	28
Collected dialogues	192	224	224
Task completion rate	95.8% (184/192)	84.8% (190/224)	79.0% (177/224)

Metrics

1. Slot accuracy	2. Insertion error rate	3. Deletion error rate
4. Substitution error rate	5. Slot error rate	6. Update precision
7. Update insertion error rate	8. Update deletion error rate	9. Update substitution error rate
10. Speech understanding rate	11. Slot accuracy for filled slots	12. Deletion error rate for filled slots
13. Substitution error rate for filled slots		

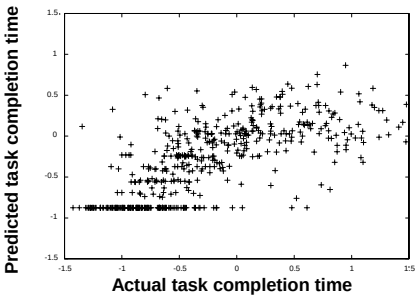
Experiment: Results

- In both domains, obtained regression models show relatively good correlation with system performance.
- Support vector regression performs better than multiple linear regression.
- The analysis of the obtained models revealed that update precision is the most important factor in reducing the task completion time.

	Multiple Linear Regression	Support Vector Regression
WI	0.488 (0.549)	0.471 (0.323)
MR-1	0.291 (0.649)	0.370 (0.367)
MR-2	0.478 (0.557)	0.494 (0.326)
MR-1+MR-2	0.432 (0.572)	0.442 (0.335)
WI+ MR-1+MR-2	0.415 (0.583)	0.456 (0.325)

R-square and
Root mean square error (in brackets)

Distribution of actual and predicted task completion times (SVR, WI+MR-1+MR-2)



Conclusion

- We proposed a method for creating an evaluation measure for discourse understanding in spoken dialogue systems.
- Obtained measures show good correlation with system performance.
- The correctness of a dialogue state update is the most important factor for system improvement.