Evaluating Discourse Understanding in Spoken Dialogue Systems

Ryuichiro Higashinaka, Noboru Miyazaki, Mikio Nakano, Kiyoaki Aikawa

NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation
3-1 Morinosato-Wakamiya, Atsugi, Kanagawa 243-0198, Japan
{rh,nmiya,nakano}@atom.brl.ntt.co.jp, aik@idea.brl.ntt.co.jp

Abstract

This paper describes a method for creating an evaluation measure for discourse understanding in spoken dialogue systems. Discourse understanding means utterance understanding taking the context into account. Since the measure needs to be determined based on its correlation with the system's performance, conventional measures, such as the concept error rate, cannot be easily applied. Using the multiple linear regression analysis, we have previously shown that the weighted sum of various metrics concerning dialogue states can be used for the evaluation of discourse understanding in a single domain. This paper reports the progress of our work: verification of our approach by additional experiments in another domain. The support vector regression method performs better than the multiple linear regression method in creating the measure, indicating non-linearity in mapping the metrics to the system's performance. The results give strong support for our approach and hint at its suitability as a universal evaluation measure for discourse understanding.

1. Introduction

Due to advances in speech recognition and speech synthesis technologies, spoken dialogue systems have been attracting a lot of attention. There are two types of spoken dialogue systems: those that understand a single user utterance and respond to it without taking context into account, and those that deal with multiple exchanges of utterances by understanding user utterances in the context of dialogues. The latter, which is discussed in this paper, has to be able to appropriately update the dialogue state each time a user utterance is made [1]. Here, a *dialogue state* means all the information that the system possesses concerning the dialogue. For example, a dialogue state includes intention recognition results after each user utterance, the user utterance history, the system utterance history, and so forth.

Although the concept error rate (CER) or the keyword error rate has been widely used as an evaluation measure in single utterance understanding, it may not be appropriate for the evaluation of discourse understanding, since it is not certain whether the CER correlates closely with the system's performance. In our previous work [2], we have shown that, in a meeting room reservation domain, the weighted sum of various metrics concerning dialogue states can evaluate discourse understanding. We defined dialogue performance by task completion time, and performed a multiple linear regression analysis using task completion time as the explained variable and the metrics as explaining variables. The obtained multiple regression model fits comparatively well and has shown its validity as an evaluation measure. However, some issues still remained. For example, currently it is not clear whether the same approach can succeed in other domains, and it is not known whether a model obtained in one domain can be applied to another. In this work, we performed additional experiments in another domain (the weather information service domain) and re-collected dialogues in the meeting room reservation domain to resolve the remaining issues. We aim at establishing a domain-independent universal evaluation measure of discourse understanding. If we have such a measure, we will be able to build spoken dialogue systems having good discourse understanding, use it for the automatic improvement of discourse understanding components, and test those components using simulations without performing costly dialogue data collection.

The next section describes the issues in detail. After that, experiments using two systems that are different in their domains are described, followed by the results. The last section summarizes and mentions future work.

2. Issues

Firstly, although we showed that discourse understanding can be evaluated by the weighted sum of the metrics concerning the dialogue states in the meeting room reservation domain, it is not clear whether the same approach can succeed in other domains. It also has to be checked whether measures obtained from several different systems in the same domain can keep consistency. Second, even if our approach is proven successful in other domains, it still would not be clear whether an evaluation measure derived from one domain is usable in the other. If it is indeed usable, the obtained measure could be considered a universal evaluation measure of discourse understanding. Thirdly, in the linear regression equation in our previous report, several coefficients did not match human intuition. For example, it seems as if the overall task completion time decreases with increasing update insertion error. Dependencies between the metrics may explain this; however, further investigation into the obtained regression model is needed. Fourthly, we previously used a stepwise multiple linear regression analysis. Since it may be difficult to linearly estimate the task completion time, it may be necessary to take into consideration other regression models to enhance this aspect. Finally, we used ten metrics to express the task completion time. However, there is no assurance that these metrics suffice. It may be necessary to introduce some other metrics concerning the dialogue states.

3. Data Collection

3.1. Systems

We created three systems and performed an experiment to study the above issues. One is in a weather information service domain (**WI**), and the other two are in a meeting room reservation domain (**MR-1**, **MR-2**). WI provides Japan-wide weather information. Users specify a prefecture or a city, a date, and an infor-

mation type (weather, temperature, precipitation) to obtain the desired information. It has a vocabulary of 853. The language model is a trigram trained from the randomly generated texts of acceptable phrases. MR-1 and MR-2 provide meeting room reservation service. Users specify a date, a room, and start and end times for the reservation. It has a vocabulary of 243. The language model is a trigram trained from the transcription obtained in the experiment of our previous report. The difference between MR-1 and MR-2 lies in their discourse understanding components. Both systems create multiple dialogue state candidates ordered by priority after each user utterance, and choose the highest ranked one as the best dialogue state. When deciding the best dialogue state, MR-1 preserves lower ranked dialogue states, whereas MR-2 discards them totally.

All three systems were developed using the spoken dialogue system toolkit WIT [3]. Their speech recognition engine is Julius [4] used with its attached acoustic model, and the speech synthesis engine is FinalFluet [5]. Each system has two switchable dialogue strategies. One is to keep accepting user utterances until it has enough information to fulfill a task or the user explicitly requests a system response. The other is to confirm each user utterance.

3.2. Experiment

Using the three systems, we collected dialogue data for analysis. The dialogue data were collected using naive users in acoustically insulated booths. Twelve subjects used WI. Each subject was given a task sheet listing what should be asked for. They were instructed to complete the tasks one by one. We prepared eight task patterns. Together with the two dialogue strategies, each subject performed 16 dialogues, for a total 192 dialogues collected. Twenty eight subjects used MR-1 and MR-2. Using four task patterns, two dialogue strategies and two systems, each performed 16 dialogues, and 448 dialogues were collected. We recorded system's utterances, start and end times of user's utterances, and dialogue states before and after the user utterance. The user's voice and system's voice were also recorded, and all user utterances were transcribed.

3.3. Metrics and System Performance

To express the dialogue states of an entire dialogue, we used ten metrics calculated by comparing the hypothesis dialogue states and the reference dialogue states as reported in [2]. We assume that a dialogue state is expressed as a frame expression, which is common in many systems [6]. A frame is a bundle of slots that consist of attribute-value pairs concerning a certain domain. Reference dialogue states are all hand-labelled. Table 1 lists the ten metrics. For example, the slot accuracy is the rate at which the slots have correct values and update precision the rate of having correct updates in slot changes. Sometimes the aim of a task is not to fill every slot but to fill some of them. To reflect such cases, we employ an additional three metrics, focusing on only the slots that have values. Table 2 lists the additional three metrics. They correspond to the slot accuracy, deletion error rate, and substitution error rate for the filled slots.

As in the previous report, we use the task completion time to express system performance. Since we are dealing with task-oriented spoken dialogue systems, this is appropriate. Task completion times are normalized using both the task pattern and the dialogue strategy, because the task completion times are severely affected by them.

Table 1: *Conventional ten metrics.*

1. slot accuracy	6. update precision
2. insertion error rate	7. update insertion error rate
3. deletion error rate	8. update deletion error rate
4. substitution error rate	9. update substitution error rate
5. slot error rate	10. speech understanding rate

Table 2: *Additional three metrics.*

11. slot accuracy for filled slots
12. deletion error rate for filled slots
13. substitution error rate for filled slots

4. Data Analysis

4.1. Recognition Accuracy and Task Completion Rate

Table 3 shows the metrics concerning speech recognition in WI and MR-1 + MR-2. MR-1 + MR-2 stands for the combined data of MR-1 and MR-2. When dialogues that took more than three minutes to complete the task are treated as failures, task completion rates for WI, MR-1 and MR-2 are 95.8% (184/192), 84.8% (190/224), and 79.0% (177/224), respectively.

4.2. Obtained Evaluation Measure

To create the discourse evaluation measure, we used only successful dialogues whose task completion times were available. We used two regression methods. One is the multiple linear regression that we used in the previous report, and the other is support vector regression (SVR). This time, for the multiple linear regression, the $m5'$ method [7, 8] was used for attribute selection instead of the greedy method. SVR is an optimization-based approach for solving machine learning regression problems based on support vector machines [9, 10, 11]. We used a polynomial kernel $K(\mathbf{x}, \mathbf{y}) = (\mathbf{x} \cdot \mathbf{y} + 1)^d$ where $d = 2$. We created regression models for each regression method using task completion time normalized by the task pattern and the dialogue strategy as the explained variable and the 13 metrics as explaining variables. Table 4 shows squared correlation coefficients (R^2) and the root mean square error ($RMSE$) for the two regression methods. These are the results of ten-fold cross validation. Most of the obtained regression models fit comparatively well and show validity as evaluation measures. According to the table, it is clear that SVR performs better than multiple linear regression. Although the R^2 of SVR is worse than that of multiple linear regression for WI, $RMSE$ is significantly lower. Therefore, from here, we only deal with models derived by SVR. Figure 1 shows the distribution of actual and predicted task completion times for the acquired models using WI + MR-1 + MR-2. The grouping of data, which seems like a horizontal line just above -1.0 in the y-axis, means that tasks with different task completion times were mapped to the same value, indicating possible differences of speaking timings among the subjects.

4.3. Commonality in Regression Models

To check whether a regression model trained from the data of one domain/system has commonality with that of the other, we calculated R^2 and $RMSE$ for every combination of models. Table 5 shows the results. Most of the R^2 values are between 0.4 and 0.5, suggesting that the model of one domain can be safely applied to that of the other. Since the performance of the model trained from WI + MR-1 + MR-2 exceeds all others, this

Table 3: *Speech recognition metrics.*

	Sent	Corr	Acc	Sub	Del	Ins	Err	S.Err
WI	1073	70.59%	66.08%	19.22%	10.19%	4.51%	33.92%	34.02%
MR-1 + MR-2	3613	78.71%	69.17%	17.07%	4.22%	9.54%	30.83%	38.47%

Table 4: *R^2 and $RMSE$ for multiple linear regression and support vector regression (SVR).*

	multiple linear regression	SVR
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)

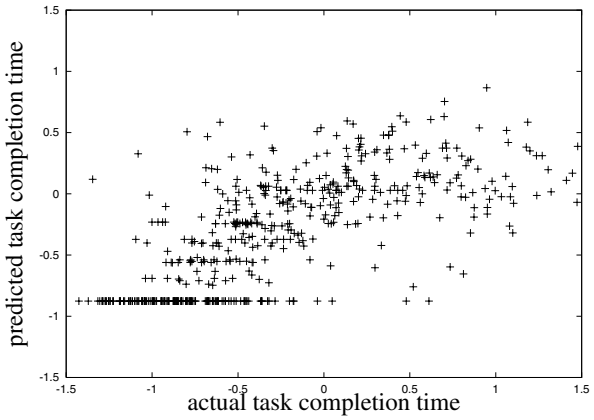

Figure 1: *Distribution of actual and predicted task completion times by the regression model trained from WI + MR-1 + MR-2.*

model could possibly be used as a universal discourse evaluation measure.

4.4. Analysis of Regression Models

Analyzing the obtained SVR models allows us to list up the possible major metrics for the prediction of the task completion time. First, the objective function of SVR is defined as

$$\begin{aligned} f(\mathbf{x}) &= \sum_{i;\mathbf{x}_i \in SVs} \alpha_i K(\mathbf{x}_i, \mathbf{x}) + b \\ &= \sum_{i;\mathbf{x}_i \in SVs} \alpha_i \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}) + b \\ &= \mathbf{w} \cdot \phi(\mathbf{x}) + b \end{aligned} \quad (1)$$

where SVs is the set of support vectors, and $\phi(\mathbf{x})$ an explicit representation of new feature vectors $\mathbf{x}$ that are mapped in the new feature space by the kernel. In the case of the 13 dimensions (features) in our original space and using second-degree polynomial kernel, the dimensions of the new feature space become 105, and $\mathbf{w}$ is written as

$$\begin{aligned} \mathbf{w} = \Big(& \sum_{i;\mathbf{x}_i \in SVs} \alpha_i x_{i1}^2, \ldots, \sum_{i;\mathbf{x}_i \in SVs} \alpha_i x_{i13}^2, \\ & \sqrt{2} \sum_{i;\mathbf{x}_i \in SVs} \alpha_i x_{i1} x_{i2}, \ldots, \sqrt{2} \sum_{i;\mathbf{x}_i \in SVs} \alpha_i x_{i12} x_{i13}, \\ & \sqrt{2} \sum_{i;\mathbf{x}_i \in SVs} \alpha_i x_{i1}, \ldots, \sqrt{2} \sum_{i;\mathbf{x}_i \in SVs} \alpha_i x_{i13}, 1 \Big) \end{aligned} \quad (2)$$

where $x_{i1} \cdots x_{i13}$ are the values of the 13 metrics (Tables 1 and 2) of the ith support vector. By gathering up the weighting factors by the metrics and the combination of the metrics, we obtain the following weights:

$$\begin{aligned} W(x_1) &= \sum_{i;\mathbf{x}_i \in SVs} \alpha_i x_{i1}^2 + \sqrt{2} \sum_{i;\mathbf{x}_i \in SVs} \alpha_i x_{i1} \\ &\vdots \\ W(x_{13}) &= \sum_{i;\mathbf{x}_i \in SVs} \alpha_i x_{i13}^2 + \sqrt{2} \sum_{i;\mathbf{x}_i \in SVs} \alpha_i x_{i13} \\ W(x_1, x_2) &= \sqrt{2} \sum_{i;\mathbf{x}_i \in SVs} \alpha_i x_{i1} x_{i2} \\ &\vdots \\ W(x_{12}, x_{13}) &= \sqrt{2} \sum_{i;\mathbf{x}_i \in SVs} \alpha_i x_{i12} x_{i13} \end{aligned}$$

We calculated all 91 weights $({}_{13}C_2 + 13)$ from the obtained models. Table 6 shows the ten most dominant metrics or the combinations of the metrics for each model with their weights. The bigger the weights are, the more significant the metrics or the combinations of the metrics become. $x_1 \cdots x_{13}$ denote the values of the 13 metrics; e.g., x_6 means the update precision, and (x_1, x_6) the combination of the slot accuracy and update precision. By a simple glance at the table, one can see that the update precision plays a key role. The speech understanding rate, denoted by x_{10} is also a dominant factor. (x_1, x_6) is also dominant in reducing the task completion time.

By comparing WI with MR-1 + MR-2, one can see that there are some differences caused by the influence of the domains. For example, x_2 (the deletion error rate) has a higher rank in WI, indicating that missing slots have a larger effect in WI than in MR. MR-1 and MR-2 have more or less the same entries in the lists, suggesting that in the same domain, dominating factors are not much different. As stated in Section 4.3, the model obtained from WI + MR-1 + MR-2 could be used as a universal evaluation measure, and we can safely say that update precision is the most important factor in reducing task completion time.

5. Summary and Future Work

This paper presented a method for creating an evaluation measure for discourse understanding in spoken dialogue systems. This paper deals with the remaining issues (Section 2) of our previous work in which we showed that the weighted sum of various metrics concerning dialogue states can be used for the evaluation in a single domain.

We collected dialogue data using three spoken dialogue systems in weather information service and meeting room reservation domains to resolve the issues. Using the multiple linear regression method and the support vector regression (SVR)

Table 5: *Commonality between the trained models.*

Training data \ Test data	WI	MR-1	MR-2	MR-1 + MR-2	WI + MR-1 + MR-2
WI	–	0.387 (0.410)	0.516 (0.324)	0.445 (0.368)	0.463 (0.344)
MR-1	0.414 (0.350)	–	0.514 (0.312)	0.449 (0.332)	0.432 (0.338)
MR-2	0.415 (0.348)	0.392 (0.385)	–	0.446 (0.350)	0.430 (0.349)
MR-1 + MR-2	0.417 (0.342)	0.401 (0.358)	0.521 (0.300)	–	0.436 (0.333)
WI + MR-1 + MR-2	0.494 (0.304)	0.423 (0.355)	0.555 (0.276)	0.481 (0.316)	–

Table 6: *Ten dominating weighting factors.*

	WI		MR-1		MR-2		MR-1 + MR-2		WI + MR-1 + MR-2	
1	$W(x_6)$	-0.456	$W(x_{10})$	-0.160	$W(x_6)$	-0.179	$W(x_6)$	-0.228	$W(x_6)$	-0.644
2	$W(x_1, x_6)$	-0.245	$W(x_6)$	-0.157	$W(x_{10})$	-0.157	$W(x_{10})$	-0.129	$W(x_1, x_6)$	-0.327
3	$W(x_2)$	0.220	$W(x_6, x_{10})$	-0.111	$W(x_8)$	0.127	$W(x_1, x_6)$	-0.127	$W(x_6, x_{11})$	-0.288
4	$W(x_6, x_{11})$	-0.184	$W(x_1, x_{10})$	-0.092	$W(x_6, x_{10})$	-0.120	$W(x_6, x_{11})$	-0.121	$W(x_8, x_{11})$	0.253
5	$W(x_8, x_{11})$	0.172	$W(x_1, x_6)$	-0.091	$W(x_1, x_6)$	-0.107	$W(x_8)$	0.117	$W(x_8, x_{10})$	0.219
6	$W(x_{12})$	-0.169	$W(x_{10}, x_{11})$	-0.090	$W(x_6, x_{11})$	-0.097	$W(x_6, x_{10})$	-0.116	$W(x_{11})$	0.195
7	$W(x_5, x_{12})$	0.164	$W(x_6, x_{11})$	-0.089	$W(x_1, x_{10})$	-0.095	$W(x_8, x_{11})$	0.084	$W(x_{12})$	-0.191
8	$W(x_9)$	0.162	$W(x_8)$	0.081	$W(x_{10}, x_{11})$	-0.090	$W(x_1, x_{10})$	-0.080	$W(x_5, x_{11})$	0.159
9	$W(x_3)$	-0.154	$W(x_1)$	-0.067	$W(x_1)$	0.085	$W(x_1)$	-0.079	$W(x_{13})$	-0.151
10	$W(x_2, x_{11})$	0.148	$W(x_8, x_{11})$	0.061	$W(x_8, x_{11})$	0.080	$W(x_1, x_8)$	0.070	$W(x_3)$	-0.150

method, we found that discourse understanding can be evaluated by the weighted sum of the metrics and the combinations of the metrics concerning the dialogue states not just in one domain, but in other domains as well, and that measures obtained from several different systems in the same domain can keep consistency among themselves. From a commonality test of the obtained regression models, it became clear that an evaluation measure derived from one domain is usable in the other, and that a universal evaluation measure of discourse understanding can be derived by creating a regression model using data from multiple domains. SVR performed better than the multiple linear regression method, suggesting a difficulty in mapping the metrics to the task completion time linearly. An analysis of the obtained regression models indicated that update precision together with slot accuracy plays the dominant role in reducing the task completion time. It should also be noted that the metrics and their effects matched human intuition. The three new metrics added this time also showed their effectiveness in that they all appeared in Table 6.

Future work includes smart handling of the horizontal line that appears in the graph and the incorporation of user-satisfaction metrics into the models.

6. Acknowledgements

We thank Dr. Hiroshi Murase and all members of the Dialogue Understanding Research Group for helpful comments. We also thank Jun Suzuki for advise on the support vector regression.

7. References

[1] V. W. Zue and J. R. Glass, "Conversational interfaces: Advances and challenges," *Proceedings of IEEE*, 88(8):1166–1180, 2000.

[2] R. Higashinaka, N. Miyazaki, M. Nakano, and K. Aikawa, "A method for evaluating incremental utterance understanding in spoken dialogue systems," *Proc. ICSLP2002*, pp. 829–832, 2002.

[3] M. Nakano, N. Miyazaki, N. Yasuda, A. Sugiyama, J. Hirasawa, K. Dohsaka, and K. Aikawa, "WIT: A toolkit for building robust and real-time spoken dialogue systems," in *Proc. SIGDIAL*, 2000, pp. 150–159.

[4] A. Lee, T. Kawahara, and K. Shikano, "Julius – an open source real-time large vocabulary recognition engine," in *Proc. Eurospeech*, 2001, pp. 1691–1694.

[5] S. Takano, K. Tanaka, H. Mizuno, M. Abe, and S. Nakajima, "A Japanese TTS Syntem Based on Multi-form Units and a Speech Modification Algorithm with Harmonics Reconstruction," *IEEE Transactions on Speech and Processing*, 9(1):3–10, 2001.

[6] D. G. Bobrow, R. M. Kaplan, M. Kay, D. A. Norman, H. Thompson, and T. Winograd, "GUS, a frame driven dialog system," *Artif. Intel.*, 8:155–173, 1977.

[7] Wang Y. and Witten I.H., "Induction of model trees for predicting continuous classes," *Proceedings of the poster papers of the European Conference on Machine Learning*, 1997.

[8] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, Morgan Kaufmann, 1999.

[9] V. N. Vapnik, *The Nature of Statistical Learnig Theory*, Springer, 1995.

[10] A. Smola and B. Sch, "A tutorial on support vector regression," *NC2-TR-1998-030*, 1998.

[11] C.-C. Chang and C.-J. Lin, *LIBSVM: a library for support vector machines*, 2001, Software available at *http://www.csie.ntu.edu.tw/˜cjlin/libsvm*.